

3/5/1 (Item 1 from file: 351)

DIALOG(R)File 351:Derwent WPI

(c) 2006 The Thomson Corporation. All rts. reserv.

0010495610 - Drawing available

WPI ACC NO: 2001-096482/200111

XRPX Acc No: N2001-073308

**Processor communication apparatus for computer network, has main memory in which erroneous logic page numbers and number of erroneous pages are stored in preset position**

Patent Assignee: NEC CORP (NIDE)

Inventor: KANO T; KANO H Y

**Patent Family** (3 patents, 2 countries)

Patent Number	Kind	Date	Application Number	Kind	Date	Update
JP 2000330960	A	20001130	JP 1999134960	A	19990514	200111 B
JP 3376956	B2	20030217	JP 1999134960	A	19990514	200316 E
US 6678722	B1	20040113	US 2000568593	A	20000511	200405 E

Priority Applications (no., kind, date): JP 1999134960 A 19990514

#### Patent Details

Number	Kind	Lan	Pg	Dwg	Filing Notes
JP 2000330960	A	JA	16	11	
JP 3376956	B2	JA	16		Previously issued patent JP 2000330960

#### Alerting Abstract JP A

NOVELTY - Reading address and write-in address of data for communication between processors (1) are designated in a logic address. The flag shows the generation of error during communication from the logic processor number of sending element of packet, which assigns communication ID to sending element. The erroneous logic page number and number of erroneous pages are stored in preset position in a main memory (4).

USE - For computer network.

ADVANTAGE - Since the erroneous page lists are stored in main memory, consideration about size of area for communication during programming is avoided.

DESCRIPTION OF DRAWINGS - The figure shows the block diagram of processor communication apparatus.

1 Processor

4 Main memory

**Title Terms/Index Terms/Additional Words:** PROCESSOR; COMMUNICATE; APPARATUS; COMPUTER; NETWORK; MAIN; MEMORY; ERROR; LOGIC; PAGE; NUMBER; STORAGE; PRESET; POSITION

#### Class Codes

International Classification (Main): G06F-015/167, G06F-015/177

(Additional/Secondary): G06F-012/00, G06F-012/08, G06F-012/10, G06F-015/00

US Classification, Issued: 709212000, 712029000, 711156000, 711202000

File Segment: EPI;

DWPI Class: T01

Manual Codes (EPI/S-X): T01-H03A; T01-H08; T01-M02C1

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-330960

(P2000-330960A)

(43) 公開日 平成12年11月30日 (2000. 11. 30)

(51) Int.Cl.<sup>7</sup>

G 0 6 F 15/177  
12/08  
12/10

識別記号

6 7 6

F I

G 0 6 F 15/177  
12/08  
12/10

テーマコード\* (参考)

6 7 6 A 5 B 0 0 5  
H 5 B 0 4 5  
F

審査請求 有 請求項の数10 O L (全 16 頁)

(21) 出願番号

特願平11-134960

(22) 出願日

平成11年5月14日 (1999. 5. 14)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 加納 健

東京都港区芝五丁目7番1号 日本電気株式会社内

(74) 代理人 100108578

弁理士 高橋 詔男 (外3名)

Fターム(参考) 5B005 JJ11 KK02 KK13 MM31 MM51

SS11 SS14

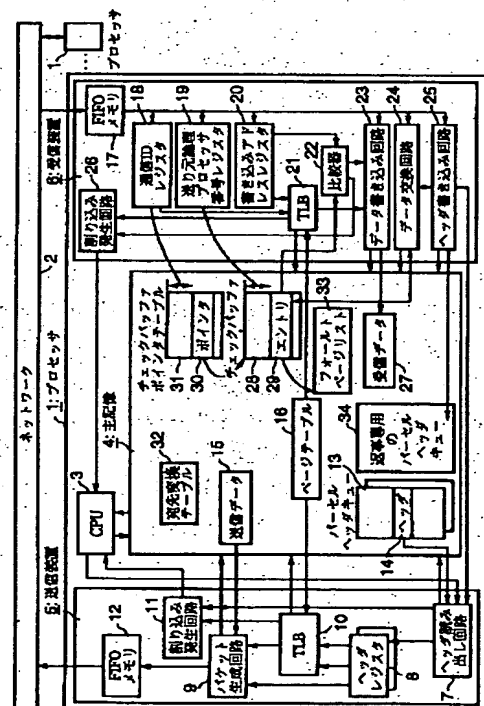
5B045 BB02 DD07 FF02 GG11

(54) 【発明の名称】 プロセッサ間通信装置

(57) 【要約】

【課題】 ページフォルトを回避するための処理によるプロセッサ間通信の性能の低下を回避し、かつ、プロセッサ間通信時のページフォルトに対応する。

【解決手段】 連続するデータのプロセッサ間通信の直後に、受信側でページフォルトが発生したかどうかを確かめるためのパケットを送る。ページフォルトが起こっていない場合には何もせず、ページフォルトが起こっていた場合には、ページフォルトした部分データだけを再送する。また、受信側で、ページフォルトしたページを記憶しておき、再送の時に使用する。さらに、ページフォルトが発生しても割り込みが頻繁に発生しないように、同じページでページフォルトしたかどうかを受信装置が確認し、同じならば割り込みを発生しない。



## 【特許請求の範囲】

【請求項 1】 主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、  
プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、  
該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備え、  
タスクごとに割り当てられた通信 ID と、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に、該送り元の論理プロセッサから該通信 ID を割り当てられたタスクに送られたプロセッサ間通信時にページフォールトが発生したかどうかを示すフラグと、該プロセッサ間通信時に直前にページフォールトした論理ページ番号と、該プロセッサ間通信時にページフォールトしたページ数と、ページフォールトした論理ページ番号を格納する主記憶上の位置を記憶する手段を備えたプロセッサ間通信装置。

【請求項 2】 前記受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合に、  
ページフォールトフラグ、該プロセッサ間通信時に直前にページフォールトした論理ページ番号、該プロセッサ間通信時にページフォールトしたページ数、ページフォールトした論理ページ番号を格納する主記憶上の位置を記憶する手段に該アドレス変換の結果に関するデータを記憶し、さらに該アドレス変換にかかる論理ページをページリストに加えることを特徴とする請求項 1 に記載のプロセッサ間通信装置。

【請求項 3】 前記受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合に、  
前記アドレス変換にかかる論理ページを予め用意しておいたダミーページを割り当ててことを特徴とする請求項 2 に記載のプロセッサ間通信装置。

【請求項 4】 主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、  
プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、  
該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備え、  
該論理アドレスを構成する論理ページの情報として、該論理ページの割り当てられた物理ページ番号と、ページの状態として下記 (i) ~ (i.v) の状態を持つことを特徴とするプロセッサ間通信装置。

- (i) 物理ページが割り当てられていない無効の状態；
- (i i) 現在主記憶になく外部記憶媒体に追い出されているページアウトの状態；
- (i i i) 以前はページアウトの状態而现在ページイン

処理を行なっている際中であるページイン中の状態；

- (i v) 該主記憶上に割り付けられているページインの状態。

【請求項 5】 主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、  
プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、  
該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備え、  
該受信装置での論理アドレスから物理アドレスへの変換時に、タスクごとに割り当てられた通信 ID と、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較する手段を持つプロセッサ間通信装置。

【請求項 6】 前記受信装置内でのアドレス変換の結果、ページの状態がページイン中の場合には、  
前記受信装置での論理アドレスから物理アドレスへの変換時に、タスクごとに割り当てられた通信 ID と、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較し、  
該論理ページ番号が同じ場合には何もせず、該論理ページ番号が異なった場合にだけ、その論理ページ番号をページリストに加え、直前にページフォールトした論理ページ番号として記憶し、ページフォールトしたページ数を 1 増やす、ことを特徴とする請求項 5 に記載のプロセッサ間通信装置。

【請求項 7】 主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、  
プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、  
該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備え、  
宛先プロセッサ内の主記憶上の、タスクごとに割り当てられた通信 ID と、パケットの送り元の論理プロセッサ番号から決定される場所のデータと、パケットに付加したデータとを交換し、自プロセッサの指定された主記憶上の位置に交換したデータを書き込む機能を持つパケットを持つプロセッサ間通信装置。

【請求項 8】 プロセッサ間通信でデータを送った直後に、前記パケットを用いて、事前に送ったプロセッサ間通信の受信側でのページフォールトがあったかどうかを知ることを特徴とする請求項 7 に記載のプロセッサ間通信装置。

【請求項 9】 プロセッサ間通信時にページフォールト

した論理ページ番号が逐次格納されたフォールトページリストを、送り元プロセッサがリモートリードを用いて読み出して、該フォールトページリストにある論理ページ番号のデータだけを再送することを特徴とする請求項8に記載のプロセッサ間通信装置。

【請求項10】 前記再送の後で、前記パケットでページフォールトの有無を再度確認し、ページフォールトが起こらなくなるまで、再送をつづけることを特徴とする請求項9に記載のプロセッサ間通信装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、複数のプロセッサがネットワークによって結合された並列コンピュータでの、仮想記憶を実現するためのプロセッサ間通信装置に関する。

【0002】

【従来の技術】複数のプロセッサが1つの問題を解く並列処理において、プロセッサ間通信は必須である。ここで、プロセッサ間通信は複数のプロセッサ間でのデータのやりとりをいう。また、このプロセッサ間通信は、複数のプロセッサで並列に処理することにより発生するのであって、並列処理のオーバヘッドの1つとなり、プロセッサ間通信を高速に行なうことが、並列処理効果を高めるためには必須となっている。

【0003】一方、並列計算機を複数のユーザで使うには、1人のユーザが処理に使用するプロセッサを決めて、1つのプロセッサでは1人のユーザのジョブだけが走るように、システムを空間的に分割するスペースシェアリングという手法が採られてきた。しかしながら、この方法では、各ユーザが使用するプロセッサ数の合計は、物理的なプロセッサ数を越えられないという制限がある。このような制限を克服して、複数のユーザが並列計算機を使う方法の1つとして、時間的にプロセッサを分割して使うタイムシェアリングの方法がある。タイムシェアリングを行なうには、複数ユーザのジョブのメモリイメージを外部の記憶媒体、例えばハードディスクに退避する仮想記憶のサポートが必須となる。

【0004】また、科学技術計算では、プロセッサ性能の向上により、それが扱うデータ規模はますます増大していく。並列処理でのデータの分配は、まず、1つのプロセッサにディスクやホストマシンからデータを転送し、分割したデータをそれぞれのプロセッサに分配する方法が採られることが多い。このような場合、分配後にはデータは各プロセッサの主記憶に格納されるが、最初に1つのプロセッサの主記憶にデータをロードする時点で、データ量が主記憶容量を越えてしまうことがある。このような問題に対応するためにも、外部の記憶媒体、例えばハードディスクを使って実主記憶容量以上のデータを扱う仮想記憶のサポートが必須である。

【0005】

【発明が解決しようとする課題】しかし、仮想記憶を並列計算機でサポートするには、次のような課題がある。即ち、プロセッサ間通信が高速化したため、ページアウトしたページをディスクから主記憶にページインする時間が相対的に非常に長くなったことである。そのため、ページインのためにプロセッサ間通信を止めた時の、他のプロセッサ間通信への影響の大きさが問題になってくる。

【0006】図1-1は、従来の並列計算機システムのページインの説明図で、(A)は全ての領域がページインされている場合、(B)は通信バッファがページインされている場合、(C)はプロセッサ間通信により送信データと受信領域がページインされている場合を示している。従来は、図1-1(A)～(C)の構成により、ページインのためにプロセッサ間通信を止めた時の、他のプロセッサ間通信への干渉を防止していた。

【0007】図1-1(A)に示す従来方法1は、仮想記憶を使わないという解である。プロセッサ間通信の対象になる主記憶領域は限定できないので、プログラムが使用する全領域を実メモリに割り当てて、ページアウトしないことにするのである。この方法では、上に述べた並列計算機での仮想記憶サポートの要求に答えられないという課題がある。

【0008】図1-1(B)に示す従来方法2は、プロセッサ間通信のための送信バッファ、受信バッファを設けるものである。この方法では、送信バッファ、受信バッファだけを常に実メモリに割り付けておき、プロセッサ間通信するデータは必ず、この送信メモリバッファと受信メモリバッファを介して転送される。そうすることにより、プロセッサ間通信自体は、実メモリから実メモリへの転送となり、プロセッサ間通信を止めることはない。しかしながら、プロセッサ間通信には、送信バッファへのデータのコピーと、受信バッファから宛先アドレスへのデータのコピーの2回のデータコピーが必要となり、プロセッサ間通信性能が著しく低下するという課題がある。

【0009】また、図1-1(C)に示す従来方法3は、通信時にそのプロセッサ間通信の送信データの領域と受信領域を予め実メモリ上に割り付けてから、プロセッサ間通信を行なうものである。この方法では、コピーすることなくデータが宛先アドレスに転送されるが、最初に実メモリ上に割り当てられていることを確認するためのプロセッサ間のやりとりが必要となる。これにより、コピーを行なわなくても、実質的なプロセッサ間通信の性能が低下する。

【0010】以上を簡潔にまとめると、従来方法1では仮想記憶自体をサポートできていない。従来方式2では、送信バッファへのデータのコピーと、受信バッファから宛先アドレスへのデータのコピーのオーバヘッドから、通信性能が著しく低下する。従来方式3では、たと

え、プロセッサ間通信に関わるメモリ領域が実メモリ領域に割り付けられていたとしても、必ず、プロセッサ間通信の前に実メモリへの割り付けを確認するためのプロセッサ間通信が必要である。そこで、この確認のためのプロセッサ間通信により、プロセッサ間通信の性能が低下する。即ち、従来の並列計算機システムで仮想記憶をサポートした場合に、実メモリ上に割り当てられた領域間のプロセッサ間通信の性能が低下してしまうという課題があった。

【0011】本発明は上述する課題を解決するもので、プロセッサ間通信に関わるメモリ領域が実メモリ領域に割り付けられている場合でも、性能低下の原因になる作業をせずに、プロセッサ間通信の性能を良好に保持するプロセッサ間通信装置を提供することを目的とする。

【0012】

【課題を解決するための手段】上記課題を解決する本発明の請求項1に記載のプロセッサ間通信装置は、図1に示すように、CPU3、主記憶4、送信装置5、並びに受信装置6を有するプロセッサ1が複数個、ネットワーク2によって結合された並列コンピュータに用いられ、プロセッサ間通信装置は、プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、送信装置5と受信装置6内に論理アドレスから物理アドレスに変換する手段TLB10、21を有している。

【0013】また、図1、図5に示すように、プロセッサ間通信装置は、タスクごとに割り当てられた通信IDとプロセッサ間通信のパケットを送った送り元の論理プロセッサ番号から決まる主記憶上の位置29に、該通信IDを宛先とした該論理プロセッサ番号のプロセッサからのプロセッサ間通信によるデータ転送中に、ページフォールトが発生したかを示すフラグ291と、該データ転送中で直近にページフォールトした論理ページ番号292と、ページフォールトしたページを記憶するページリストの先頭アドレスを記憶する手段293と、ページフォールトしたページ数とを記憶する手段294を持つ構成である。

【0014】好ましくは、本発明のプロセッサ間通信装置では、受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合には、前記のページフォールトに関するデータを通信IDと送り元論理プロセッサ番号によって決まる主記憶上の位置に記憶し、アドレス変換にかかる論理ページを主記憶上のフォールトページリスト33に加える構成とするとよい。さらに好ましくは、図8のステップS110に示すように、アドレス変換にかかる論理ページを予め用意しておいたダミーページを割り当てる構成とするとよい。

【0015】また、本発明の請求項4に記載のプロセッサ間通信装置では、図4に示すように、TLB10は、論理アドレス201を構成する論理ページ202の情報

として、論理ページの割り当てられた物理ページ番号229と、ページの状態230として下記(i)～(iv)の状態を持つことを特徴とするものである。

(i) 物理ページが割り当てられていない無効の状態；  
(ii) 現在主記憶になく外部記憶媒体に追い出されているページアウトの状態；

(iii) 以前はページアウトの状態であるページイン処理を行なっている際中であるページイン中の状態；

(iv) 該主記憶上に割り付けられているページインの状態。

【0016】また、本発明の請求項5に記載のプロセッサ間通信装置は、図4に示すように、受信装置でのアドレス変換時に、タスクごとに割り当てられた通信IDと、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較する手段222、223を持つ構成としている。

【0017】さらに、好ましくは、本発明のプロセッサ間通信装置では、図8に示すように、受信装置内でのアドレス変換の結果、ページの状態230がページイン中の場合には、比較器22を用いて論理ページの比較を行なう。そして、論理ページ番号が同じ場合には何もせず、論理ページ番号が異なった場合にだけ、その論理ページ番号をフォールトページリスト33に加え、直近にページフォールトした論理ページ番号292として記憶し、ページフォールトしたページ数294を1増やす、ことを行なう構成とするとよい。

【0018】また、本発明の請求項7に記載のプロセッサ間通信装置では、図3に示すように、宛先の通信IDと送り元の論理プロセッサ番号から決まる主記憶上の位置のデータと、パケットに付加されたデータとを交換し、交換した結果を送り元のプロセッサの決められた主記憶上の位置に書き込むチェックパケット133を持つ。

【0019】好ましくは、本発明のプロセッサ間通信装置では、プロセッサ間通信でデータを送った直後に、チェックパケットを用いて、事前に送ったプロセッサ間通信の受信側でのページフォールトがあったかどうかを知る構成とするとよい。

【0020】さらに、本発明のプロセッサ間通信装置では、第2の実施の形態で示すように、作成されるページフォールトしたページのリストを、リモートリードによって、送り元のプロセッサが読み出して、そこにある論理ページ番号のデータだけを再送し、再度前記パケットでページフォールトの有無を確認し、ページフォールトが起こらなくなるまで、再送をつづける構成としている。

【0021】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態を説明する。図1は、本発明の一実施の形態を

10

20

30

40

50

説明する構成ブロック図である。図において、並列コンピュータは、複数のプロセッサ1がネットワーク2に接続されて構成されている。プロセッサ1は、CPU3と主記憶4と送信装置5と受信装置6から構成される。送信装置5内は、ヘッダ読み出し回路7、ヘッダレジスタ8、パケット生成回路9、TLB10、割り込み発生回路11、並びにFIFOメモリ12から構成されている。

【0022】ヘッダ読み出し回路7は、CPU3からの送信要求によって、主記憶4上に設けられたパーセルヘッダキュー13からヘッダ14を読み出し、ヘッダの内容を検査し、エラーがなければ、ヘッダレジスタ8に格納する。ヘッダレジスタ8は、複数のヘッダ14を格納できる。

【0023】パケット生成回路9は、ヘッダレジスタ8内の複数のヘッダ14のうち次にパケットを送信できるものを選び、そのヘッダに書かれたソースアドレスをTLB10を使って論理アドレスから物理アドレスに変換する。そして、ヘッダレジスタ8内のヘッダ14を使って、ネットワーク2に送り出すパケットのヘッダを作成し、TLB10により変換された物理ソースアドレスからのデータの読み出しを主記憶4に依頼し、読み出した送信データ15をパケットとしてネットワーク2に送り出す。

【0024】TLB10は、ヘッダレジスタ8内の通信IDとソースアドレスから、論理アドレスを物理アドレスに変換する。TLB10内にはアドレス変換に必要な情報がキャッシュされているが、必要に応じて、主記憶4上のページテーブル16にアクセスしたり、OSにページの割り当てを要求するための割り込みをおこしたりして、アドレス変換に必要な情報を得る。

【0025】FIFOメモリ12は、ネットワーク2とパケット生成回路9間に設けられたもので、データ幅の変換やエラー検出符号などの作成を行なう。割り込み発生回路11は、ヘッダ読み出し回路7でのヘッダの検査の結果、エラーとなった場合や、TLB10における論理アドレスから物理アドレスへの変換の際に、OSに対して処理を依頼しなければならない時に、CPU3に対して割り込みを発生させる。

【0026】受信装置6内は、FIFOメモリ17、通信IDレジスタ18、送り元論理プロセッサ番号レジスタ19、書き込みアドレスレジスタ20、TLB21、比較器22、データ書き込み回路23、データ交換回路24、ヘッダ書き込み回路25、並びに割り込み発生回路26から構成される。

【0027】FIFOメモリ17は、ネットワークから送られてきたパケットを一次的に格納する。通信IDレジスタ18は、パケットヘッダに含まれる通信IDを格納する。送り元論理プロセッサ番号レジスタ19は、パケットヘッダに含まれる、そのパケットを送信したプロ

セッサの論理プロセッサ番号を格納する。書き込みアドレスレジスタ20は、パケットヘッダに含まれている、パケットのデータの書き込みアドレス（デスティネーションアドレス）を格納する。

【0028】TLB21は、通信IDレジスタ18内の通信IDと、書き込みアドレスレジスタ20内のデスティネーションアドレスから、論理アドレスを物理アドレスに変換して、データ書き込み回路23に物理書き込みアドレスを渡す。TLB21内にはアドレス変換に必要な情報がキャッシュされているが、必要に応じて、主記憶4上のページテーブル16にアクセスしたり、OSにページの割り当てを要求するための割り込みをおこしたりして、アドレス変換に必要な情報を得る。

【0029】データ書き込み回路23は、パケットで送られてきた受信データ27を、主記憶4上の、TLB21から出力された書き込みアドレスに、書き込む。データ交換回路24は、主記憶4上に設けられたチェックバッファ28のエントリ29（以下、チェックバッファエントリ29という）の値と、パケットで送られてきたデータを交換し、チェックバッファエントリ29の値をヘッダ書き込み回路25に出力する。

【0030】ヘッダ書き込み回路25は、送られてきたパケットに含まれる情報と、データ交換回路24からのデータとから、受信したパケットの返事のパーセルヘッダを作成し、主記憶4上の返事専用のパーセルヘッダキュー34にヘッダを書き込み、送信装置5内のヘッダ読み出し回路7に送信要求を出す。

【0031】比較器22は、書き込みアドレスレジスタ20に格納された書き込みアドレスの一部と、主記憶4上のチェックバッファエントリ29の一部を比較し、比較結果をデータ書き込み回路23に出力し、必要に応じて割り込み発生回路26を使って割り込みを発生させる。

【0032】割り込み発生回路26は、比較器22での比較結果で割り込みが必要な場合や、TLB21における論理アドレスから物理アドレスへの変換の際に、OSに対して処理を依頼しなければならない時に、CPU3に対して割り込みを発生させる。

【0033】チェックバッファポインタテーブル31は主記憶4に設けられたもので、各通信IDのチェックバッファ28へのポインタ30を格納している。受信装置6は、通信IDレジスタ18内の通信IDと、主記憶4上のチェックバッファポインタテーブル31を使って、その通信IDに対応するチェックバッファ28をアクセスする。

【0034】宛先変換テーブル32は、主記憶4に設けられたもので、送信装置5内のヘッダ読み出し回路7が、宛先論理プロセッサ番号を、宛先物理プロセッサ番号と通信IDに変換するのに用いられる。フォールトページリスト33は、主記憶4に設けられたもので、ペー

ジフォールトしたページの論理ページ番号を格納する。

【0035】次に、この並列コンピュータで通信時に使われる、パーセルとパケットについて説明する。パーセルは通信の単位であって、そのヘッダ14がパーセルヘッダキュー13に書かれる。ヘッダレジスタ8には、ヘッダ読み出し回路7が、主記憶4上の宛先変換テーブル32を使って、このパーセルヘッダの宛先を論理プロセッサ番号から、物理プロセッサ番号と通信IDに、変換したものが書き込まれる。パケットはネットワーク内を流れる通信の単位であり、ネットワーク2を構成するスイッチ内のFIFOメモリの容量等により、その大きさが限定される。1つのパーセルで送られるデータはパケット生成回路9により、複数のパケットに分割され送られる。

【0036】図2は、この並列コンピュータで通信されるリモートDMAパーセルの構成図で、(A)はパーセルヘッダキュー13内での形式、(B)はヘッダレジスタ8内での形式、(C)はリモートDMAパケットの形式を示している。

【0037】まず、リモートDMAパーセルは、送信データの読み出しアドレス105からデータ長104分のデータを、宛先論理プロセッサ番号103で示された宛先のプロセッサ1の主記憶4の書き込みアドレス106へ書き込む機能を持つ。リモートDMAのパーセルヘッダ101の先頭にはコード欄102が設けられている。コード欄102に書き込まれた「000」は、このパーセルヘッダがリモートDMAのヘッダであることを示している。

【0038】ヘッダ読み出し回路7により、主記憶4上のパーセルヘッダキュー13から読み出されたリモートDMAのパーセルヘッダ101は宛先変換テーブル32を使って、宛先論理プロセッサ番号103が、宛先物理プロセッサ番号と通信ID110に変換されて、ヘッダレジスタ8に書き込まれる。

【0039】パケット生成回路9は、ヘッダレジスタ8内のリモートDMAのパーセルヘッダ108と、そのタスクが持っている情報121を使って、リモートDMAのパケット114を生成する。リモートDMAのパケット114は、リモートDMAを示す「000」が書き込まれたコード欄115と、宛先の物理プロセッサ番号と通信ID116と、パケット長117、送り元の論理プロセッサ番号118、書き込みアドレス119、送信データ120から構成される。

【0040】パケット生成回路9は、ヘッダレジスタ8内の残りデータ長111と、送信データの読み出しアドレス112と書き込みアドレス113からパケット長117を決定し、リモートDMAのパケット114を生成する。ヘッダレジスタ8内のリモートDMAのパーセルヘッダの残りデータ長111と、読み出しアドレス112、書き込みアドレス113は、1つのパケットを送信

するごとに更新され、パケット生成回路9は、残りデータ長111がゼロになるまで、パケットを送る。

【0041】パケット生成回路9は、1つのパケットの送信や受信の途中にページフォールト等が起こらないように、読み出しアドレス112と書き込みアドレス113のページ境界をまたぐパケットは生成しない。即ち、パケット生成回路9は、読み出しアドレス112のページ境界と書き込みアドレス113のページ境界では、必ず、パケットが切られるように、パケット長117を決定する。

【0042】次にチェックパケットのパーセルについて説明する。図3は、この並列コンピュータで通信されるチェックパケットのパーセルとチェックパケットの返事のパーセルの構成図で、(A)はパーセルヘッダキュー13内での形式、(B)はヘッダレジスタ8内での形式、(C)はチェックパケットの形式、(D)はパーセルヘッダキュー13内での形式、(E)はヘッダレジスタ8内での形式、(F)はチェックパケットの返事のパケットの形式を示している。チェックパケットのパーセルは、特別なアドレスにある値と交換するデータ126を交換し、値をチェックパケットの返事のパケット152として返送し、指定された返事の書き込みアドレス125に書き込む機能を持つ。

【0043】パーセルヘッダキュー内のチェックパケットのパーセルヘッダ122には、チェックパケットのパーセルであることを示す「100」が書き込まれたコード欄123、宛先の論理プロセッサ番号124、チェックパケットの返事を書き込む書き込みアドレス125、交換するデータ126から構成される。チェックパケットのパーセルヘッダ122もヘッダレジスタ8に書き込まれる前に、宛先変換テーブル32を使って、宛先論理プロセッサ番号124が宛先の物理プロセッサ番号と通信ID130に変換される。

【0044】パケット生成回路9は、ヘッダレジスタ内のチェックパケットのパーセルヘッダ128と、そのタスクの情報141から、次のようなものから構成されるチェックパケット133を作成する。コード欄134に書き込まれた「100」は、パケットの種類がチェックパケットであることを示す。送り元の論理プロセッサ番号137と送り元の物理プロセッサ番号と通信ID138は、そのタスクが持っている情報141である。返事の書き込みアドレス139と交換するデータ140は、チェックパケットのパーセルヘッダ128のものをそのまま使う。チェックパケットのパーセルヘッダ128は必ず1つのチェックパケット133として送信される。

【0045】最後にチェックパケットの返事のパーセルヘッダ142について、図3を参照して説明する。このパーセルヘッダは受信装置6のデータ交換回路24で交換されたチェックバッファエントリ29と、チェックパケット133内にある情報とから作成され、ヘッダ書き

込み回路25によって主記憶4上の返事専用のパーセルヘッダキュー34に書き込まれる。

【0046】チェックパケットの返事のパーセルヘッダ142には、チェックパケットの返事を示す「101」が書き込まれたコード欄143と、チェックパケット133の送り元の物理プロセッサ番号と通信ID138をそのまま使った宛先の物理プロセッサ番号と通信ID144、チェックパケット133の返事の書き込みアドレス139をそのまま使った返事の書き込みアドレス145と、交換したチェックバッファエントリ146から構成される。

【0047】ヘッダ読み出し回路7は、返事専用のパーセルヘッダキューから、チェックパケットの返事のパーセルヘッダ142を読み出すと、チェックパケットの返事を示す「101」が書き込まれたコード欄143を確認する。他のリモートDMAのパーセルヘッダ101やチェックパケットのパーセルヘッダ122の場合とは異なり、ヘッダ読み出し回路7は、宛先プロセッサ番号を変換しないで、パーセルヘッダキュー内のチェックパケットの返事のパーセルヘッダ142に書かれた宛先の物理プロセッサ番号と通信ID144をそのままヘッダレジスタ8に書き込む。これは、チェックパケットの返事のパーセルヘッダ142が、送信装置5の packets 生成回路9によって作られた送り元の物理プロセッサ番号と通信ID138をもとに、受信装置6のヘッダ書き込み回路25によって、作られたものであり、ユーザタスクがアクセスできない特別なパーセルヘッダキュー34に書かれたものであるため、そのまま使っても安全であるからである。

【0048】パケット生成回路9は、チェックパケットの返事のパーセルヘッダ147からチェックパケットの返事の packets 152を作成し、ネットワークに送り出す。チェックパケットの返事のパーセル147も、必ず1つのチェックパケットの返事の packets 141となる。

【0049】次にTLB10、21で行なわれるアドレス変換について図4を使って説明する。図4はTLB10の詳細を説明する構成ブロック図である。なお、TLB21もTLB10と同様の構成となっている。図4において、論理アドレス201は、論理ページ番号202とオフセット203によって構成されている。TLB10では、通信ID204と論理ページ番号202から、物理ページ番号229とページの状態230を引く。この実施例では、TLB10には4個のエントリがあり、与えられた通信ID204と論理ページ番号202と、4個のエントリ内の通信ID206~209と論理ページ番号210~213とが比較器222、223によって比較され、両方が同じものの物理ページ番号214~217とページの状態218~221が選択回路224からの信号により、MUX225、226によって、選

ばれ、そのエントリの物理ページ番号229とページの状態230が出力される。

【0050】もし、4個のエントリ内に、通信IDと論理ページ番号の組が与えられた通信ID204と論理ページ番号202に等しいものがなければ、TLB10は主記憶4上のページテーブル16をアクセスして、該当するページ情報をTLB10内の4個のエントリのうちの1個に読み出す。もし、ページテーブル16にもそのページ情報がなければ、TLB10はOSに割り込みをあげて新しいページ情報を設定してもらう。

【0051】主記憶4上のページテーブル16の構成はいろいろなものが考えられる。最も簡単なものは、通信ID204と論理ページ番号202を組み合わせたものをオフセットとする構成である。主記憶4の使用効率等を考えると、論理アドレス201を論理ページ番号202とオフセット203だけではなく、セグメントと論理ページ番号とオフセットと解釈し、セグメントごとにページテーブル16を用意する構成としてもよい。

【0052】ただし、ページの状態230として、次のような状態を定義する必要がある。ページの状態230は2bitで表されており、次のようなものがある。

【0053】“00”：そのページが無効であることを示す。新規にアクセスされたページの場合もあるので、割り込みを発生させて、OSに物理ページの割り付けを依頼する。

【0054】“01”：そのページが主記憶4になくディスクにページアウトされている状態であることを示す。この場合も割り込みを発生させて、OSに対し処理を依頼する。

【0055】“10”：そのページが主記憶4になくディスクにページアウトされているが、ページイン中であることを示す。この場合は、割り込みを発生させない。

【0056】“11”：そのページが主記憶4上にページインされていることを示す。

【0057】次にチェックバッファエントリ29について説明する。図5に示すように、チェックバッファエントリ29には、ページフォールトが起こったかどうかを示すページフォールトフラグ291、直前にページフォールトした論理ページ番号292、ページフォールトしたページ数294、フォールトページリスト33の先頭アドレス293が格納されている。ページフォールトフラグ291は、“0”の時にページフォールトが起こっていないことを、“1”の時に、ページフォールトがすでに起こったことを示す。

【0058】次にフォールトページリスト33について説明する。フォールトページリスト33は、1つのチェックバッファエントリ29に対して1つ設けられており、その先頭アドレス293はチェックバッファエントリ29に書かれている。フォールトページリスト33には、ページフォールトした論理ページ番号331が書き



込まれている。フォールトページリスト33の大きさは、チェックバッファエントリ29内のページフォールトしたページ数294となる。1つのパーセルヘッダで送信できるデータの大きさは限られているので、フォールトページリスト33の大きさの上限は、「1パーセルで転送できるデータ」÷「1ページの大きさ」となる。

【0059】次に本発明でのプロセッサ間通信の方法に関して説明する。まず、前述したように、パーセルが通信の単位である。1つのリモートDMAパーセルでは限られた大きさの連続したアドレスのデータが通信できる。本実施例では、1つのリモートDMAパーセルで送れるデータの大きさは512kバイトとする。また、1ページの大きさは4kバイトとする。従って、フォールトページリスト33の大きさは最大で128ページ分となる。このくらいの大きさなら、フォールトページリスト33の領域を仮想記憶のページアウトの対象から外し、必ずページインの状態に固定することも可能である。

【0060】本発明では、1つのリモートDMAパーセルを送るごとに、その直後にチェックパケットのパーセルを送る。

【0061】図6に、通信ライブラリの処理を示す。まず、通信ライブラリは、送信するデータを1パーセルで送れる大きさにN分割する(ステップS201)。次に、Nが“0”かどうか調べ(ステップS202)、“0”でないなら、1個のリモートDMAパーセルを送る(ステップS203)。その直後に1個のチェックパケットのパーセルを送る(ステップS204)。チェックパケットで送る交換するデータは、ページフォールトフラグ291が“0”のチェックバッファエントリ29の初期値となる。

【0062】そして、そのチェックパケットの返事を待つ(ステップS205)。チェックパケットの返事は、直前に送ったリモートDMAパーセルのデータ受信時にページフォールトが起こったかどうかを示すチェックバッファエントリ29である。そのチェックバッファエントリ29のページフォールトフラグ291を調べる(ステップS206)。そして、もし、ページフォールトが起こっていれば、その処理を行なって、このリモートDMAパーセルの送信を完了する(ステップS207)。

【0063】ページフォールトに対する処理は、要するに、このリモートDMAパーセルで送ったデータが正確に受信されるように処理するもので、詳細は後で説明する。一方、もし、ページフォールトが起こっていなければ、何もしない。そして、Nを1減らし(ステップS208)、Nがゼロになるまで、繰り返す。

【0064】次に、図1、図2、及び図7を使って送信時のページフォールトに対する処理について説明する。送信装置5内のパケット生成回路9は、ヘッダレジスタ8内のパーセルヘッダを1つ選び(ステップS00

1)、そのパケットを生成する。そのパーセルヘッダがリモートDMAのパーセルヘッダである場合(ステップS002)には、パケット生成回路9は、その送信データ読み出しアドレスをTLB10を使って物理アドレスに変換する(ステップS003)。そしてページの状態230を調べる(ステップS004)。

【0065】ページの状態230が無効“00”なら、TLB10は割り込みを発生させて(ステップS005)、OSに新しいページを割り当ててもらう(ステップS006)。通常は転送するデータは事前にCPUが作成しているので、読み出しアドレスのページの状態230が“00”であることはほとんどない。そして、OSはページの状態230を“11”にして(ステップS007)、TLB10を割り込み状態から復帰する(ステップS008)。TLB10は、また、ページの状態230を調べる(ステップS004)。

【0066】ページの状態230がページアウト“01”なら、TLB10は割り込みを発生させて(ステップS009)、OSにページインの操作を開始してもらう(ステップS010)。さらにOSは、ページの状態230を“10”に書き換えページイン中の状態にする(ステップS011)。そして、OSはTLB10を割り込み状態から復帰させる(ステップS012)。TLB10は、また、ページの状態230を調べる(ステップS004)。

【0067】ページの状態230がページイン中“10”なら、TLB10は割り込みを発生させず、ページの状態230をパケット生成回路9に伝える。パケット生成回路9は、そのパケットのデータがページイン中であるので、そのパーセルヘッダのパケットの送信を中止し(ステップS013)、他のパーセルヘッダの送信を試みるため、パーセルヘッダを選ぶ(ステップS001)。

【0068】ページの状態230がページイン“11”なら、TLB10は割り込みを発生させず、ページの状態230をパケット生成回路9に伝える。パケット生成回路9は、主記憶4の送信データ15をアクセスし、このパケットを送信する(ステップS014)。そして、パケット長をもとに、ヘッダレジスタ内の残りデータ長、読み出しアドレス、書き込みアドレスを更新する(ステップS015)。そして、送ったパケットが最後かどうか、つまり、残りデータ長がゼロかどうかを調べる(ステップS016)。

【0069】もし、最後のパケットでなければ、次のパケットを送るためパーセルヘッダの選択を行なう(ステップS001)。もし、最後のパケットの場合には、ヘッダレジスタ8からこのパーセルヘッダを消して、ヘッダレジスタ8を解放する(ステップS017)。そして、次に送るパーセルヘッダを選択する(ステップS001)。

【0070】パケット生成回路9が選んだパーセルヘッダが、リモートDMAのパーセルではなかった場合、つまり、チェックパケットのパーセルヘッダか、チェックパケットの返事のパーセルヘッダの場合には、TLB10を使ってアドレスの変換をする必要はなく、そのままパケットを送信する。これらのパーセルは1個のパケットにしかならないので、パーセルヘッダをヘッダレジスタ8から消し、ヘッダレジスタを解放する(ステップS017)。そして、次のパーセルヘッダを選択する(ステップS001)。

【0071】次に、図1、図5、及び図8を使って、受信側でのページフォールトに対する処理について説明する。受信装置6は、リモートDMAのパケットが到着すると、パケットヘッダの情報を、通信IDレジスタ18、送り元プロセッサ番号レジスタ19、書き込みアドレスレジスタ20に格納する(ステップS101)。そして、通信IDレジスタ18内の通信IDと、書き込みアドレスレジスタ20内の書き込みアドレスから、TLB21を使って、論理アドレスを物理アドレスに変換する(ステップS102)。

【0072】TLB21は、入力された通信IDと論理ページ番号から、そのページの物理ページ番号とページの状態230を引く。そして、そのページの状態230を調べる(ステップS103)。

【0073】ページの状態230が無効“00”なら、TLB21は割り込みを発生させて、OSに新しいページの割り当てを行なってもらふ(ステップS104)。OSは、主記憶上にある新しいページをその論理ページに割り当てて物理ページ番号を設定し(ステップS105)、ページの状態230を“11”にし(ステップS106)、割り込み処理から抜ける(ステップS107)。TLB21は、新たに設定されたページ情報を調べる(ステップS103)。

【0074】ページの状態230がページアウト“01”なら、TLB21は割り込みを発生させて、OSに以下のような処理を依頼する(ステップS108)。まず、OSはそのページのページインの処理を開始する(ステップS109)。そして、OSは受信データを書き捨てるために予め用意していたダミーページをその物理ページ番号に設定し(ステップS110)、ページの状態230をページイン中“10”にする(ステップS111)。

【0075】次にOSは、このパケットの通信IDと送り元の論理プロセッサ番号から、該当するチェックバッファエントリ29をアクセスし、ページフォールトフラグ291を調べる(ステップS120)。もし、以前にページフォールトが起こっていない場合には、ページフォールトフラグ291を“1”にして、フォールトページリスト33の先頭アドレスをチェックバッファエントリ29に設定する(ステップS121)。

【0076】そして、チェックバッファエントリ29にこの論理ページ番号を、直近にページフォールトした論理ページ番号292として書き込み(ステップS122)、チェックバッファエントリ29のページフォールトしたページ数を1増やし(ステップS123)、フォールトページリスト33にこの論理ページ番号を加える(ステップS124)。そして、割り込み処理から復帰する(ステップS125)。TLB21は、新たに設定されたページ情報を調べる(ステップS103)。

10 【0077】ページの状態230がページイン中“10”なら、TLB21は割り込みを発生させず、ページの状態230を受信装置6に伝える。受信装置6は、そのパケットの通信IDからチェックバッファポインタテーブル31をアクセスし、その通信IDのチェックバッファの先頭アドレス30を取り出し、送り元論理プロセッサ番号レジスタ19内のそのパケットの送り元の論理プロセッサ番号をオフセットとして、チェックバッファエントリ29をアクセスする。そして、比較器22を使って、そこに書かれた論理ページ番号と、このページの論理ページ番号とを、比較する(ステップS113)。

20 【0078】比較の結果、2つのページ番号が同じならば、データ書き込み回路23は、その物理アドレスに受信データを書き込む。この物理アドレスはダミーのページであり、送られてきたデータは捨てられることになる(ステップS117)。比較の結果、2つのページ番号が異なれば、比較器22は割り込みを発生させて(ステップS115)、OSに次のことを依頼する。

30 【0079】OSは、このパケットの通信IDと送り元の論理プロセッサ番号から、該当するチェックバッファエントリ29をアクセスし、ページフォールトフラグ291を調べる(ステップS120)。もし、以前にページフォールトが起こっていない場合には、ページフォールトフラグ291を“1”にして、フォールトページリスト33の先頭アドレスをチェックバッファエントリ29に設定する(ステップS121)。

40 【0080】そして、チェックバッファエントリ29にこの論理ページ番号を、直近にページフォールトした論理ページ番号292として書き込み(ステップS122)、チェックバッファエントリ29のページフォールトしたページ数を1増やし(ステップS123)、フォールトページリスト33にこの論理ページ番号を加える(ステップS124)。そして、割り込み処理から復帰する(ステップS125)。そして、データ書き込み回路23は、その物理アドレスに受信データを書き込む。この物理アドレスはダミーのページであり、送られてきたデータは捨てられることになる(ステップS117)。

50 【0081】ページの状態230がページイン“11”なら、TLB21は割り込みを発生させず、ページ情報をデータ書き込み回路23に伝える。データ書き込み回

路23は、主記憶にアクセスして、送られてきたデータを書き込む(ステップS118)。このパケットは正常に受信されたことになる。

【0082】最後に、受信時にページフォールトが起こった場合の送信側での通信ライブラリの処理について説明する。チェックパケットのパーセルを送って、その返事を調べた時、ページフォールトフラグ291が“1”であった場合には、ページフォールトが起こったということである。通信ライブラリは、割り込みをおこして、OSに対し、ページフォールトしたページへのデータの転送を依頼し、スリープする。

【0083】OSは、受信側のOSと通信を行ない、チェックバッファエンドリ29に書かれたページフォールトしたページ数と、フォールトページリスト33の先頭アドレスを使って、ダミーページに書き込まれて捨てられたデータを再送する。そして、リモートDMAのパーセルで送ろうとしたデータが全部受信された時点で、スリープしていた通信ライブラリをおこす。

【0084】第2の実施の形態では、新たにリモートリードパーセルを導入する。図9は、リモートリードパーセルとチェックパケットの返事のパーセルの構成図で、(A)はパーセルヘッダキュー13内での形式、(B)はヘッダレジスタ8内での形式、(C)はリモートリードパケットの形式、(D)はパーセルヘッダキュー13内での形式、(E)はヘッダレジスタ8内での形式、(F)はリモートリードの返事のパケットの形式を示している。

【0085】リモートリードパーセルは、図9に示すように、宛先論理プロセッサ番号403で指定されたプロセッサの読み出しアドレス405から、データ長404分のデータを読み出し、リモートリードパーセルを送ったプロセッサの主記憶4の書き込みアドレス406に書き込む機能を持つ。コード欄402に書き込まれた「110」は、これがリモートリードのパーセルヘッダ401であることを示す。リモートリードのパーセルヘッダ401もヘッダレジスタ8に書き込まれる前に、宛先変換テーブル32を使って、宛先論理プロセッサ番号403が宛先の物理プロセッサ番号と通信ID410に変換される。

【0086】パケット生成回路9は、ヘッダレジスタ内のリモートリードのパーセルヘッダ408と、そのタスクの情報422から、次のようなものから構成されるリモートリードパケット414を作成する。「110」が書き込まれたコード欄415は、それがリモートリードのパケットであることを示す。送り元の物理プロセッサ番号と通信ID418は、そのタスクが持っている情報141である。データ長419と読み出しアドレス420、書き込みアドレス421は、リモートリードのパーセルヘッダ408のものをそのまま使う。リモートリードのパーセルヘッダ408は必ず1つのリモートリード

のパケット414として送信される。

【0087】次に、リモートリードの返事のパーセルヘッダ423について説明する。このパーセルヘッダは受信装置6のヘッダ書き込み回路25によって、リモートリードパケット414内にある情報から作成され、主記憶4上の返事専用のパーセルヘッダキュー34に書き込まれる。

【0088】リモートリードの返事のパーセルヘッダ423には、リモートリードの返事を示す「111」が書き込まれたコード欄424と、リモートリードパケット414の送り元の物理プロセッサ番号と通信ID418をそのまま使った宛先の物理プロセッサ番号と通信ID425、データ長426、読み出しアドレス427、書き込みアドレス428から構成される。

【0089】ヘッダ読み出し回路7は、返事専用のパーセルヘッダキュー34から、リモートリードの返事のパーセルヘッダ423を読み出すと、リモートリードの返事を示す「111」が書き込まれたコード欄424を確認し、チェックパケットの返事のパーセルヘッダの場合と同じように、宛先プロセッサ番号を変換しないで、パーセルヘッダキュー内のリモートリードの返事のパーセルヘッダ423に書かれた宛先の物理プロセッサ番号と通信ID425をそのままヘッダレジスタ8に書き込む。

【0090】これは、リモートリードの返事のパーセルヘッダ142が、送信装置のパケット生成回路9によって作られた送り元の物理プロセッサ番号と通信ID418をもとに、受信装置6のヘッダ書き込み回路25によって、作られたものであり、ユーザタスクがアクセスできない特別なパーセルヘッダキュー34に書かれたものであるため、そのまま使っても安全であるからである。

【0091】パケット生成回路9は、リモートDMAのパーセルを処理するのと同じように、残りデータ長432と、読み出しアドレス433、書き込みアドレス434から、パケット長438を決定し、リモートリードの返事のパケット435を送り出す。リモートリードの返事のパケットは、複数送られることがある。

【0092】リモートリードの返事のパーセルはリモートDMAのパーセルと同じように、送信装置5で処理される。つまり、送信アドレスをTLB10で論理アドレスから物理アドレスに変換しないといけないので、送信データがページアウトしている可能性もある。リモートリードの返事のパーセルは、図7でのリモートDMAのパーセルの送信処理と同じく処理される。従って、リモートリードの返事のパーセルを送る時の送信側のページフォールトの問題は適切に処理される。

【0093】第2の実施の形態では、このリモートリードパーセルを使って、フォールトページリスト33を送信側通信ライブラリが読み、該当するページの部分を再送するものである。

【0094】図1.0は通信ライブラリの処理を説明する図で、ここでは適宜図5も参照して説明する。通信ライブラリは、1個のリモートDMAのパーセルを送った(ステップS301)後に、チェックパケットを送る(ステップS302)。そして、そのチェックパケットの返事を受信(ステップS303)し、チェックバッファエントリ29のページフォルトフラグ29.1を調べる(ステップS304)。もし、ページフォルトが起こっていなければ、次のパーセルの送信を行なう。

【0095】もし、ページフォルトが発生していたら、送られてきたチェックバッファエントリ29内にある、フォルトページリスト33の先頭アドレスから、ページフォルトしたページ分のデータをリモートリードパーセルによって読み出す(ステップS305)。そして、フォルトページリスト33に書かれた論理ページに送ったデータを再送する(ステップS306)。

【0096】この時、ページフォルトしたページは連続していないかもしれないが、フォルトページリスト33に書かれている順番、つまり、論理ページ番号が大きくなる順番に、そのデータを複数のパーセルで送る。そして、最後に、まだ、チェックパケットを送信してページフォルトの有無を確認する(ステップS302)。

【0097】通信ライブラリは、ページフォルトしたページがすでにページインされていることを期待して再送するが、まだ、ページインされていないかもしれない。もし、ページインされていないページがあれば、また、受信側でダミーページに書き捨てられることになる。通信ライブラリは、ページフォルトしたページだけのデータを再送し、ページフォルトしたページはいつかはページインされるので、いつかは、再送でのページフォルトが発生しなくなる。

【0098】第2の実施の形態によると、OSを介さずに、ユーザタスクが再送をする構成としているので、OSの負担を軽減できるという効果がある。また、フォルトページリスト33に関しては、前述したように、あまり大きなリストにはならないので、事前にページイン状態に固定した領域を使うことで、フォルトページリスト33そのもののページフォルトを避けることができる。

【0099】さらに、たとえフォルトページリスト33がページアウトされていても、リモートリードの返事のパーセルは、図7で示したリモートDMAのパーセルと同じく処理されるので、送信データに関するページフォルトは適切に処理される。一方、リモートリードの返事のパケットの受信時のページフォルトであるが、フォルトページリスト33の大きさは、たかだか1ページ分であることと、リモートリードしている側が受信側なので、受信するページをCPUがアクセスすることで、事前にページインできることから、ページフォール

トは回避できると考えられる。

【0100】

【発明の効果】以上説明したように、請求項1に記載のプロセッサ間通信装置によれば、プロセッサ間通信に使われる領域の大きさを意識しないでプログラミングできることである。なぜなら、チェックバッファエントリとフォルトページリストは主記憶に記憶されるが、プログラムで使われる領域はすべて仮想記憶の対象なので、従来方式3のようにプロセッサ間通信する主記憶の領域を制限する必要がないからである。

【0101】請求項3に記載のプロセッサ間通信装置によれば、ページフォルトが発生しても、ネットワークの閉塞が起こらないという効果がある。なぜなら、ページフォルトしたページへのデータの書き込みは、通常のデータ受信動作と変わらず、ダミーページに書き込むことで、捨てられるからである。

【0102】請求項4に記載のプロセッサ間通信装置によれば、プロセッサ間通信の性能を低下させずにすむことである。なぜなら、論理ページの状態として、無効、ページアウト、ページイン中、及びページインの4状態に分けて、ページフォルトの確認を円滑に行なうからである。

【0103】請求項5に記載のプロセッサ間通信装置によれば、ページフォルトが発生した場合でも受信装置が発生する割り込みが少ないことである。なぜなら、チェックバッファエントリ内の直近にページフォルトした論理ページ番号と、今ページフォルトした論理ページ番号を受信装置が比較するため、同じページのページフォルトによる複数の割り込みを発生させずにすむからである。

【0104】請求項7及び請求項8に記載のプロセッサ間通信装置によれば、ページフォルトが発生しない場合には、プロセッサ間通信の性能を低下させずにすむことである。なぜなら、ページフォルトの確認は大量のデータを送った後に1度だけ、しかも、ページフォルトしたことを確認するためのパケットの処理を受信装置が高速に行なうからである。

【0105】請求項9に記載のプロセッサ間通信装置によれば、ページフォルトした時にOSの負担を軽減できるという効果がある。なぜなら、第2の実施の形態で導入したリモートリードパーセルを使うことにより、ユーザタスクがフォルトページリストを読み出して、自分でページフォルトで受信されなかったデータを再送することができるからである。

【図面の簡単な説明】

【図1】 本発明のプロセッサ間通信装置の構成図である。

【図2】 この並列コンピュータで通信されるリモートDMAパーセルの構成図である。

【図3】 この並列コンピュータで通信されるチェック

パケットのパーセルとチェックパケットの返事のパーセルの構成図である。

【図4】 本発明の実施の形態で用いるTLBの構成図である。

【図5】 本発明で用いるチェックバッファエントリとフォールトページリストの構成図である。

【図6】 本発明での送信側の通信ライブラリの処理を説明するフローチャートである。

【図7】 本発明での送信側の読み出しアドレスがページフォールトした場合の送信装置での処理について説明するフローチャートである。

【図8】 本発明での受信側の書き込みアドレスがページフォールトした場合の受信装置での処理について説明するフローチャートである。

【図9】 本発明の第2の実施の形態で導入するリモートリードのパーセルとパケットの構成図である。

【図10】 本発明の第2の実施の形態での、受信側でページフォールトが起こった場合の、通信ライブラリの処理について説明するフローチャートである。

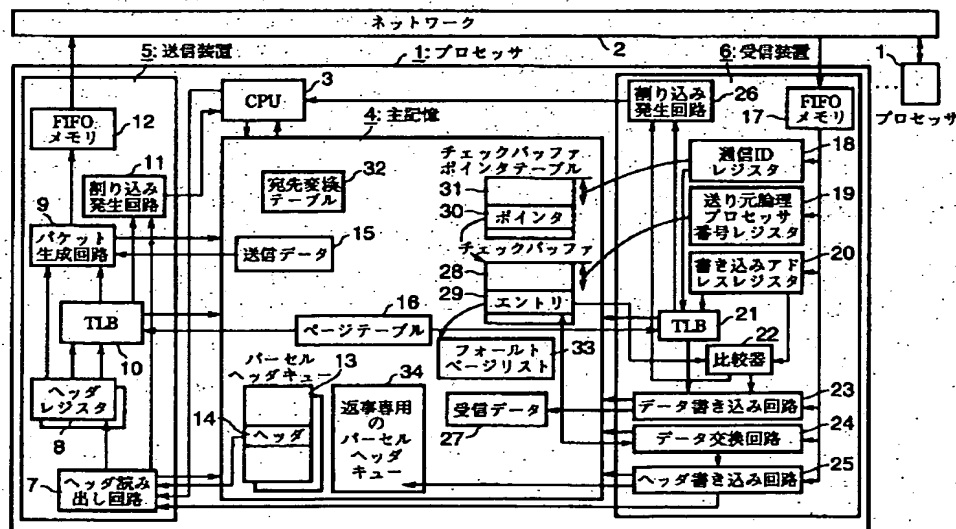
【図11】 従来の並列計算機システムのページインの説明図である。

【符号の説明】

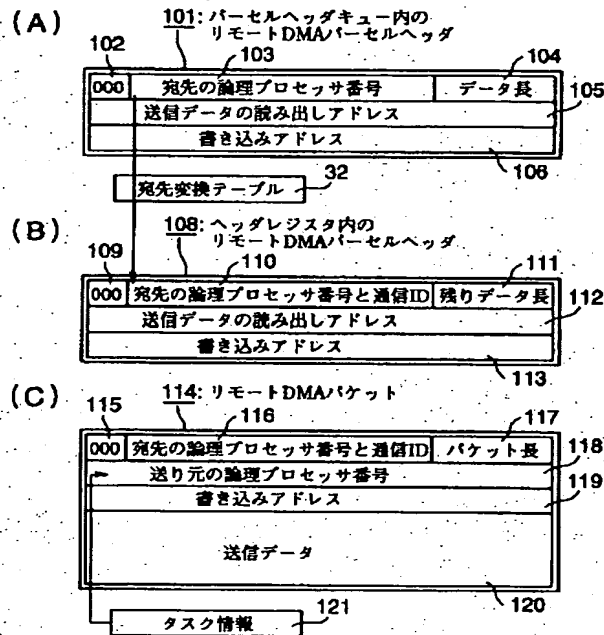
- 1 プロセッサ
- 2 ネットワーク
- 3 CPU
- 4 主記憶
- 5 送信装置
- 6 受信装置
- 7 ヘッダ読み出し回路

- \* 8 ヘッダレジスタ
- 9 パケット生成回路
- 10、21 TLB
- 11 割り込み発生回路
- 12 FIFOメモリ
- 13 パーセルヘッダキュー
- 14 ヘッダ
- 15 送信データ
- 16 ページテーブル
- 17 FIFOメモリ
- 18 通信IDレジスタ
- 19 送り元論理プロセッサ番号レジスタ
- 20 書き込みアドレスレジスタ
- 22 比較器
- 23 データ書き込み回路
- 24 データ交換回路
- 25 ヘッダ書き込み回路
- 26 割り込み発生回路
- 27 受信データ
- 28 チェックバッファ
- 29 チェックバッファエントリ
- 31 チェックバッファポインタテーブル
- 32 宛先変換テーブル
- 33 フォールトページリスト
- 114 リモートDMAパケット
- 133 チェックパケット
- 152 チェックパケットの返事のパケット
- 414 リモートリードパケット
- \* 435 リモートリードの返事のパケット

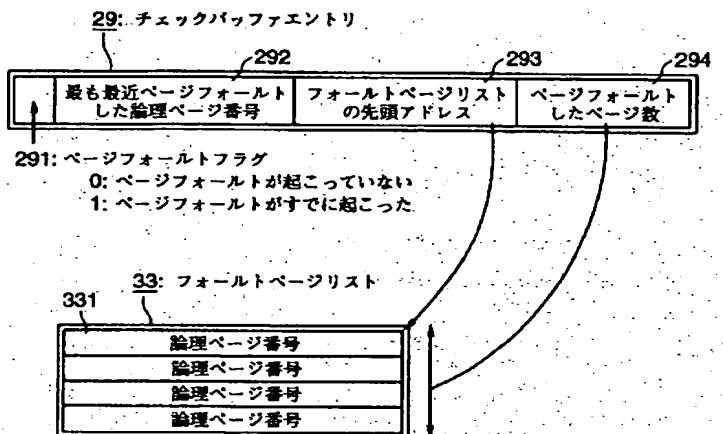
【図1】



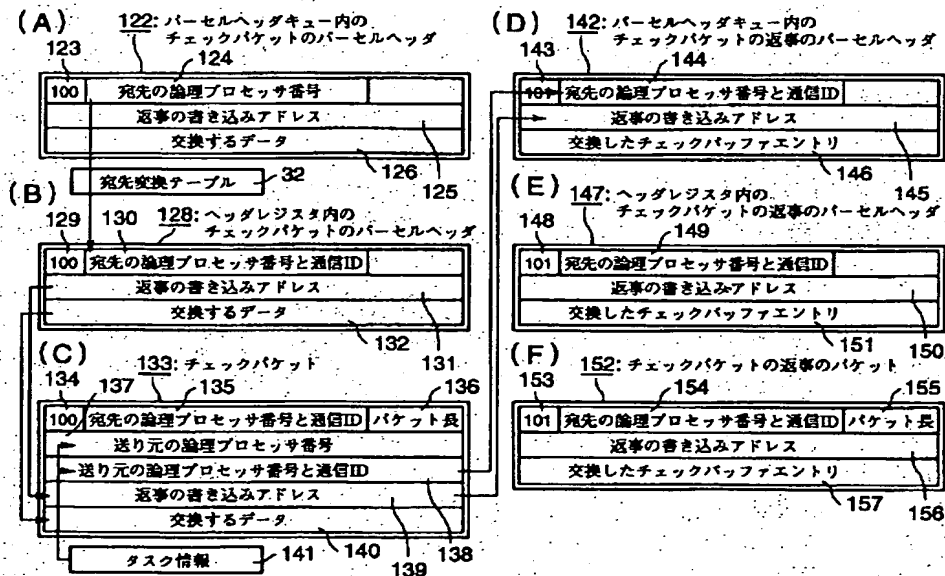
【図2】



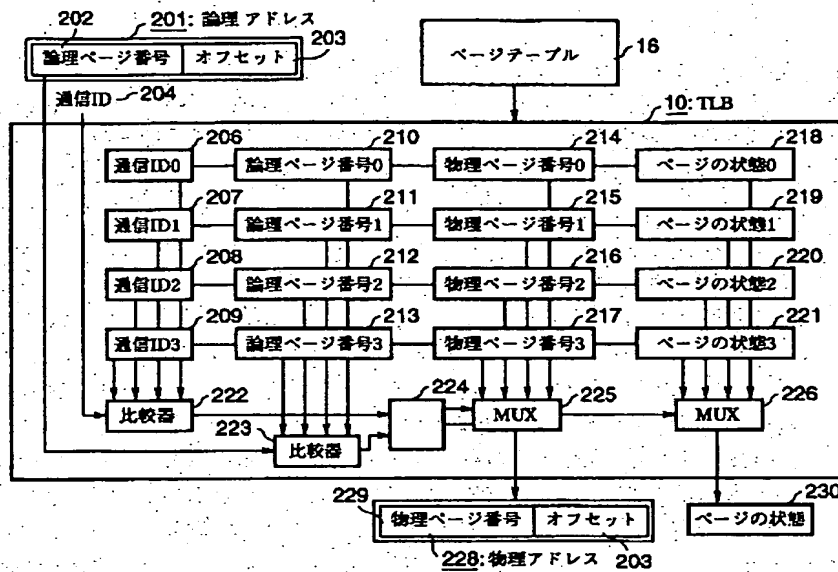
【図5】



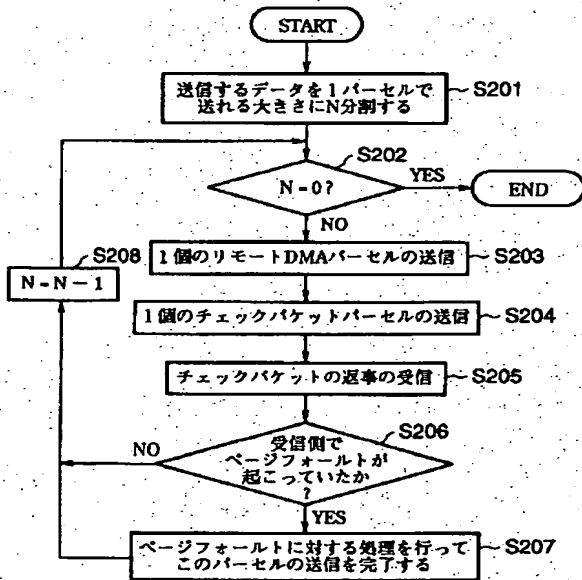
【図3】



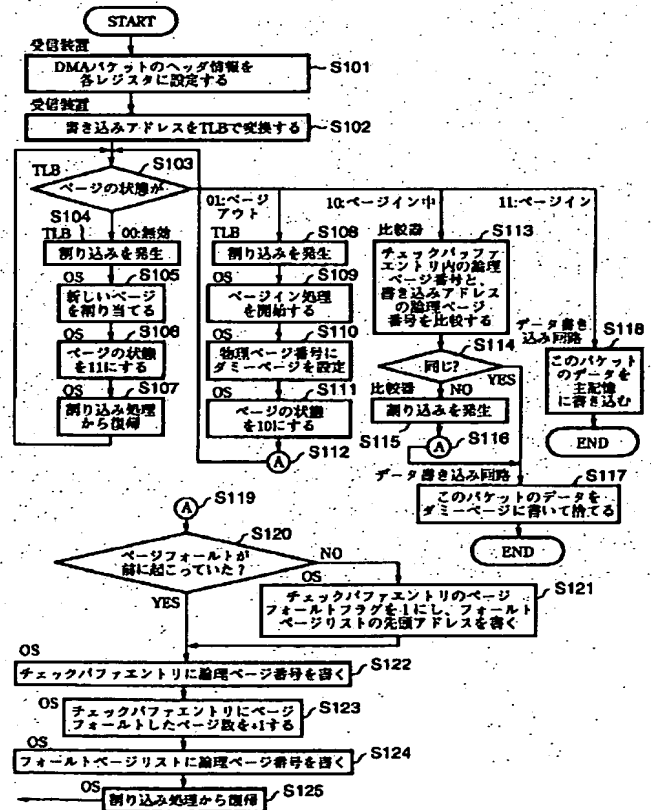
【図4】



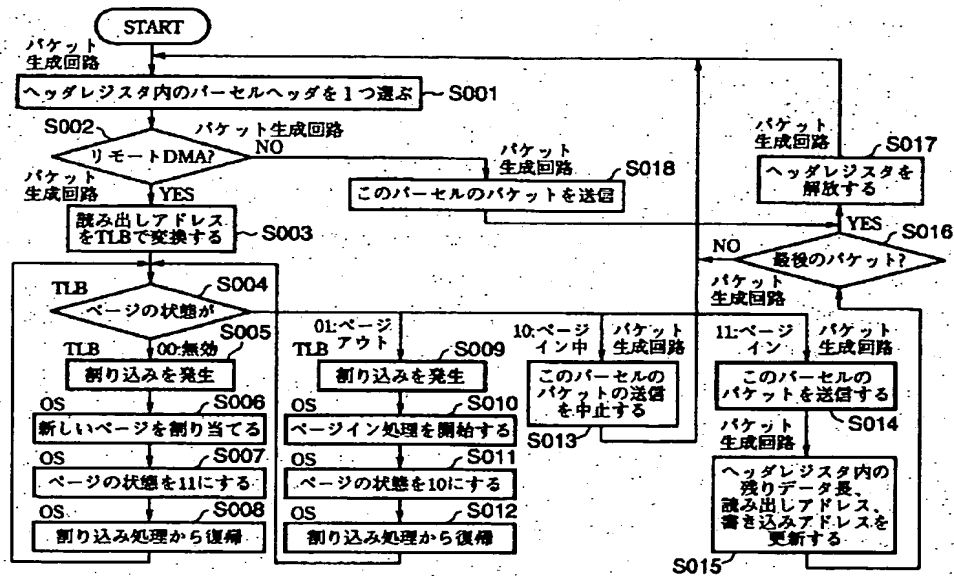
【図6】



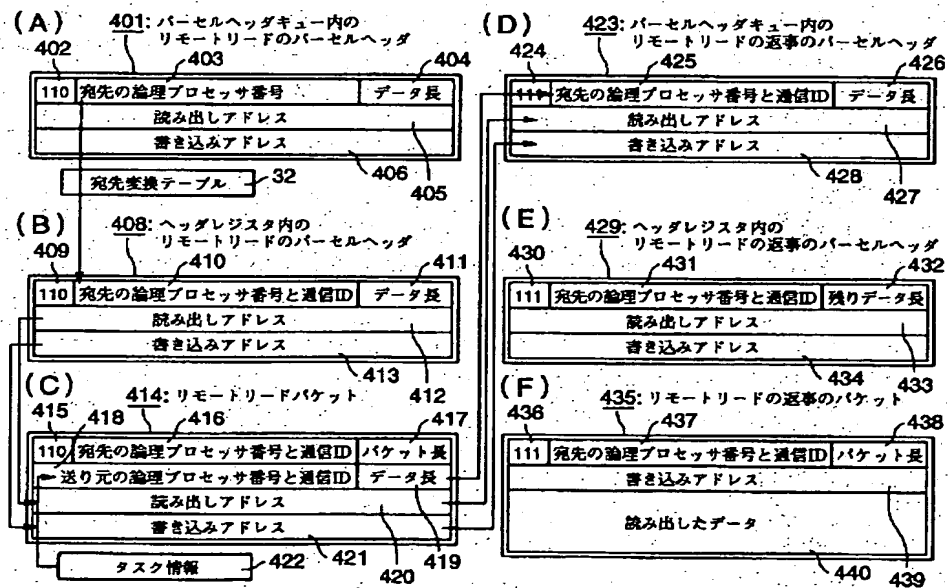
【図8】



【図7】

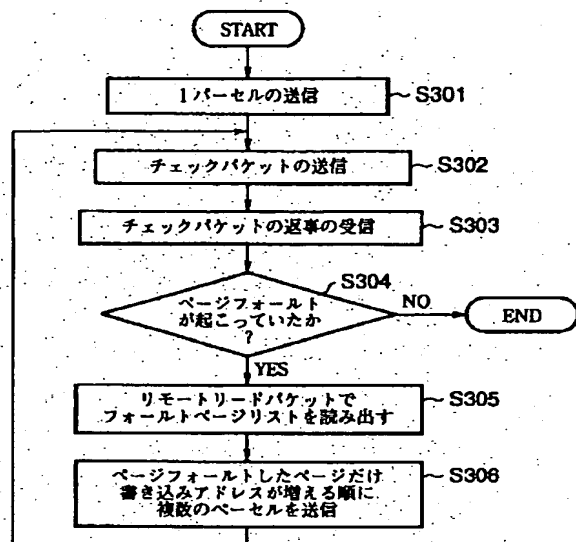


【図9】





【図10】



【図11】

